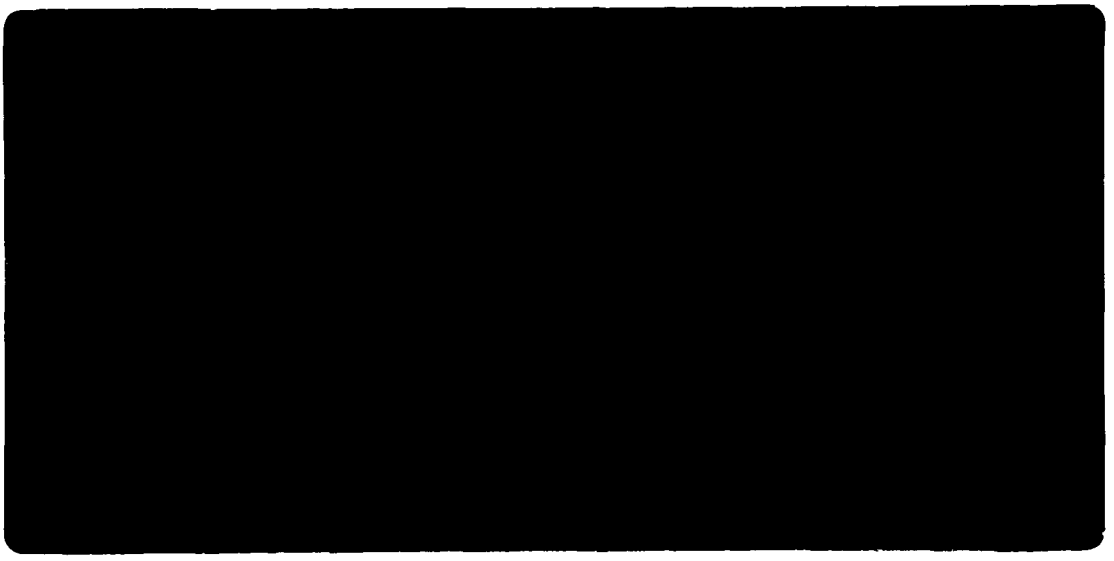
Institute of Paper Science and Technology
Atlanta, Georgia

IPST TECHNICAL PAPER SERIES



NUMBER 400

A NEW SOLUTION FOR THE PROBABILITY OF COMPLETING SETS IN RANDOM SAMPLING: DISCOVERY OF THE "TWO-DIMENSIONAL FACTORIAL"

J.D. LINDSAY

SEPTEMBER, 1991

A New Solution for the Probability of Completing Sets in Random Sampling: Discovery of the "Two-Dimensional Factorial"

J.D. Lindsay

Submitted for publication in the Journal of the American Statistical Association

A NEW SOLUTION FOR THE PROBABILITY OF COMPLETING SETS IN RANDOM SAMPLING: DISCOVERY OF THE "TWO-DIMENSIONAL FACTORIAL"

Jeffrey D. Lindsay
The Institute of Paper Science and Technology
Atlanta, GA 30318

ABSTRACT

A new solution to a classical sampling problem is found. The problem concerns the probability of completing a subset of unique items randomly distributed among an infinite population (or randomly sampled with replacement from a finite population) with each item having an equal probability of being sampled in any trial. In deriving the solution, an interesting recursive function is obtained which can be described as a "two-dimensional factorial." This function is partially tabulated, and several of its properties are investigated, including limits for large numbers. Use of this function offers significant computational advantages over the previous classical solution to the probability problem considered here. The function is not known to have been discovered in previous work.

The solution of the sampling problem is extended to the case of nonuniform probabilities for different classes of items. The two-dimensional factorial function is utilized again in this more complex solution.

Applications are discussed, and several sample calculations are offered.

INTRODUCTION

In probability theory, a classical sampling problem concerns the likelihood of collecting a set of items by randomly sampling a population[1]. A simple example can be found in the collection of sets of promotional items offered inside cereal boxes. The items are presumably randomly and uniformly distributed, and remain unidentified until the package has been opened. For instance, one cereal manufacturer offered miniature license plates from all 50 states with one plate per box. If somebody desires to collect all 50, how many boxes should one plan to purchase to be 95% confident that the set will be completed? A less ambitious consumer may simply want to know the probability that at least 10 different plates will be obtained by purchasing 12 boxes. More complex cases may be found in sampling problems in scientific studies. For example, consider the case of a paper fiber analyst microscopically characterizing individual fibers sampled from a commercial pulp that may contain many species of trees. How many individual fibers should the analyst examine to be 95% confident that at least one fiber from each species is represented?

We will begin by considering the simple problem when the different classes in the population each compose an equal fraction of the population. In general terms, our problem statement becomes:

> *If U unique classes of items are randomly and uniformly distributed among an infinite population, what is the probability that a specified number, U-M, of the unique items will be acquired in N trials? (M is the number of missing classes in the sample.)*

We will introduce the notation P(N,U-M) to denote this probability. Feller[2] shows that this probability is

$$P(N,U\text{-}M) = \binom{U}{M} \sum_{k=0}^{U-M} (-1)^k \binom{U-M}{k} \left[1 - \frac{M+k}{U}\right]^N. \qquad (1)$$

By taking an independent approach in the solution, we will show a new form for the solution to be

$$P(N,U\text{-}M) = \frac{U!}{M!\, U^N} F(D, U\text{-}M), \qquad (2)$$

where D is the number of duplicate items among the N samples, or D = N-(U-M), and F is a recursive function defined by

$$F(D,U\text{-}M) \equiv \sum_{j=1}^{U-M} j\, F(D\text{-}1,j), \qquad (3)$$

$$F(0,j) = 1 \text{ for all } j = 1,2,3,....$$

Following derivation of Equation (2) and a discussion of its relation to Equation (1), an extension will be derived for the case of nonuniform probabilities. Specifically, we will find the probability of collecting all U classes in N trials when the classes no longer have equal probabilities of being sampled. Though the solution for the nonuniform case is more complex, it is also of more practical value for practical sampling problems.

After derivation of the probability formulas, we will discuss properties and limiting values of the recursive function F, which can be described as a two-dimensional factorial function.

DERIVATION FOR UNIFORM DISTRIBUTIONS

The two-dimensional factorial function was found by noting obvious patterns while determining the permutations for obtaining U-M unique items in N trials. That number, divided by the total number of possible permutations, U^N, gives the desired probability. For example, consider the problem of collecting all three items of a set in six tries. Here U = 3, M = 0, and N = 6, and the number of duplicates, D, is 3. The permutations are treated in the following table. There are

10 cases to consider, depending on when the duplicates are encountered. Duplicates are shown in bold, italic text. For example, in case 6, duplicates occur at trials 2, 5, and 6. For trial 1, any of the three unique items can be chosen. If a duplicate is then to occur in trial 2, there is only one possibility, the same item that was selected in the first trial. Trials 2 and 3 are to be unique items, so the number of possibilities becomes 2 and 1, respectively. For trials 5 and 6, any selection will be a duplicate, so the number of possibilities becomes 3 and 3.

Case	Trials						Permutations
1.	3	***1***	***1***	***1***	2	1	=3! * (1*1*1)
2.	3	***1***	***1***	2	***2***	1	=3! * (1*1*2)
3.	3	***1***	***1***	2	1	***3***	=3! * (1*1*3)
4.	3	***1***	2	***2***	***2***	1	=3! * (1*2*2)
5.	3	***1***	2	***2***	1	***3***	=3! * (1*2*3)
6.	3	***1***	2	1	***3***	***3***	=3! * (1*3*3)
7.	3	2	***2***	***2***	***2***	1	=3! * (2*2*2)
8.	3	2	***2***	***2***	1	***3***	=3! * (2*2*3)
9.	3	2	***2***	1	***3***	***3***	=3! * (2*3*3)
10.	3	2	1	***3***	***3***	***3***	=3! * (3*3*3)
						Total:	=3!* 90 =3!*F(3,3)

Table 1. Permutations for the 10 possible cases when U=3, N = 6, and M = 0.

The total number of permutations is the product of 3! and the total permutations for duplicates, which is the sum of the products in parentheses in the rightmost column of Table 1. The sum of numbers in parentheses can be written as either

$$F(3,3) = \sum_{a_3=1}^{3} \prod a_1\, a_2\, a_3 \text{ , with } a_3 \geq a_2 \geq a_1 \text{ , } a_i \in \{1,2,3\} \qquad (4)$$

or as

$$3*(1*1 + 1*2 + 2*2 + 1*3 + 2*3 + 3*3) + 2*(1*1+1*2+2*2) + 1*(1*1)$$

$$= F(3,3) = \sum_{j\,=\,1}^{3} j\, F(2,j) \text{ ,} \qquad (5)$$

$$\text{where } F(2,j) = \sum_{a_2=1}^{j} \prod a_1\, a_2 \text{ , with } a_2 \geq a_1 \text{ , } a_i \in \{1,2 \ldots j\}. \qquad (6)$$

In general, the number of cases is given by the number of ways D duplicates can be distributed among N=U-M+D samples, with duplicates able to occur only after at least one element of U has been selected. The number of cases is thus (N-1)!/(D! [N-D-1]!). The number of choices available for a duplicate equals the number of unique items previously selected in that case.

By considering the trends in Table 1 and the restrictions imposed on the permutations for duplicate items as a function of location in each series of trials, it can be shown that the number of permutations for obtaining all U unique items in N trials, resulting in D = N-U duplicates, is

$$U! \sum_{a_D=1}^{U} \prod a_1 a_2 \ldots a_D \text{ , with } a_D \geq a_{D-1} \geq a_{D-2} \geq \ldots \geq a_1, \; a_i \in \{1,2 \ldots U\} \tag{7}$$

which can also be written as

$$U!\, F(D,U) \tag{8}$$

$$\text{where } F(D,U) \equiv \sum_{j=1}^{U} j\, F(D-1,j), \tag{9}$$

$$\text{and } F(0,j) = 1 \text{ for all } j = 1,2,3,....$$

When M of the U unique items are missing in the sampled subset, the number of duplicates becomes D = N-(U-M). By considering the permutations of duplicates and unique items, similar to what has been shown in Table 1, it is easily shown that the total number of permutations becomes

$$\frac{U!}{M!} F(D,U-M) \tag{10}$$

with the function F the same as defined in Equation (9). In general, then, the probability of obtaining U-M unique items from a possible U items, distributed uniformly throughout an infinite population, in N trials is

$$P(N,U-M) = \frac{U!}{M!\, U^N} F(D, U-M) \tag{11}$$

where D is the number of duplicate items among the N samples, D = N-(U-M), and F is a recursive function defined by

$$F(D,U-M) \equiv \sum_{j=1}^{U-M} j\, F(D-1,j), \tag{12}$$

$$F(0,j) = 1 \text{ for all } j = 1,2,3,....$$

Equating the r.h.s. of Equations (1) and (11) and simplifying yields

$$F(D,U\text{-}M) = \sum_{j=1}^{U-M} j\, F(D\text{-}1,j) = \sum_{j=0}^{U-M} \frac{(-1)^j\,(U\text{-}M\text{-}j)^N}{j!\,(U\text{-}M\text{-}j)!} \qquad (13)$$

The identity in Equation (13) is by no means obvious and is an interesting result of itself.

The probabililty P(N,U-M) can be computed using either Equation (11) or Equation (1) from Feller. Likewise, F(D,U-M) can be determined using the recursive approach of Equation (12), or the alternating-sign series in Equation (13). Use of the recursive function offers a significant computational advantage, for it is a summation of positive terms only, whereas the alternating-sign series involves small differences of large numbers. Limited numerical resolution on a computer thus greatly restricts the usefulness of Equation (1). For example, to compute F(D= 4, U = 43, M = 0) = 8.04E+11 with the alternating-sign series, differences between numbers 16 orders or magnitude greater are required. From j=6 to 11, the terms of the series are 1.45E+27, -2.04E+27, 2.35E+27, -2.25E+27, 1.80E+27, and -1.22E+27. Summing the series on a computer with 15 digits of resolution (the Wingz™ spreadsheet by Informix was used on a Macintosh II) yielded a negative result, whereas accuracy was maintained with the recursive approach until sums exceeded the largest allowed number, 1.7E+308.

NONUNIFORM DISTRIBUTIONS

The above results can be extended to nonuniform distributions among categories, but the probability p_i of sampling from any category U_i must be expressed as the ratio of an integer to a common integral denominator, W. We then treat the problem as if there were a finite series of W elements belonging to U different categories, with any category U_i having m_i members, and $\Sigma m_i = W$. If we sample with replacement or form a population from an infinite number of identical series, each item randomly sampled has a probability $p_i = m_i/W$ of belonging to category U_i.

We now seek the probability of representing all U classes in N>U samples. This probability will be termed $P(N,U,m_1,m_2,\ldots m_U)$, reflecting the integral weighting factors, m_i, for the various categories. We will treat the W items in the series as if they each form a distinct class, W_i, terming the W items as secondary classes and the U original classes as primary classes. Note that for every primary class U_i, there are m_i classes in the secondary series which belong to U_i. We now examine how the U primary categories can be filled through collecting subsets of the W secondary categories. In N trials, we may obtain all U primary categories by obtaining any of the following:

- U distinct secondary classes (of W possible) if each secondary class belongs to a distinct primary class,
- U+1 distinct secondary classes if U of the U+1 secondary classes belong to a distinct primary class, or

⋮

- min(W,N) distinct secondary classes if U of the min(W,N) sampled secondary classes belong to a distinct primary class.

For example, suppose we have three primary categories, U_1, U_2, and U_3, with probabilities of being sampled of 1/2, 1/3, and 1/6, respectively. We treat this as a series of six items, with $m_1 = 3$, $m_2 = 2$, and $m_3 = 1$. The six items of the series can be labeled as secondary categories A,B,C,D,E, and F, with A,B, and C belonging to primary category U_1, D and E belonging to U_2, and F belonging to U_3. If we sample with replacement N = 3 times, the combinations which fill all three primary classes are: (A,D,F), (A,E,F), (B,D,F), (B,E,F), (C,D,F), and (C,E,F). Each of these six sets can be selected in 3! ways, and the total number of ways to select three items is 6^3. The probability of filling all three primary classes in this case is thus $6*3!/6^3 = 1/6$.

For j = 0,1,2,...min(N,W)-U, then if we obtain U+j distinct secondary categories out of W possible in N tries, the number of missing secondary classes is M = W - (U+j) and the number of secondary duplicates is D = N - (U+j). The probability of obtaining U+j secondary categories, based on the analysis presented above, is

$$P(N,W-M) = \frac{W!}{M!\, W^N} F(D, W-M) . \tag{14}$$

We now must determine the probability that a randomly selected set of U+j secondary categories among W fills all U primary categories, which probability we term Z(U,U+j,W). We define A_i as the event that primary category U_i is missing and define its probability $Pr\{A_i\}$ as p_i. For combinations of events (an event being the missing of a primary class), we define

$$p_i = Pr\{A_i\},\ p_{i,j} = Pr\{A_iA_j\},\ p_{i,j,k} = Pr\{A_iA_jA_k\}, ... \tag{15}$$

where the subscripts are never equal and are written in increasing order for uniqueness. For example, $p_{2,5,7}$ is the probability that U_2, U_5, and U_7 are missing. The sum of all p's with r subscripts is defined as S_r, or

$$S_1 = \Sigma\, p_i,\ S_2 = \Sigma\, p_{i,j} ,\ S_3 = \Sigma\, p_{i,j,k} , ... \tag{16}$$

Each unique combination only appears once in any summation since we require i < j < k < ....U. The probability that none of the primary classes are missing is

$$Z(U,U+j,W) = 1 - P_1, \tag{17}$$

where P_1 is the probability of missing at least one missing primary class, and is given by

$$P_1 = S_1 - S_2 + S_3 - S_4 + - ... \pm S_u. \tag{18}$$

We have applied the logic and terminology of Feller[3] in obtaining this result.

We must now determine the individual p's for missing primary classes. To miss one primary class, U_i, m_i of the M missing secondary classes must belong to U_i. The probability that the M missing secondary classes are so distributed is

$$p_i = \frac{\binom{W-m_i}{M-x_i}}{\binom{W}{M}} \tag{19}$$

In general,

$$p_{i,j,\ldots q} = \frac{\binom{W-[m_i+m_j+\ldots m_q]}{M-[m_i+m_j+\ldots m_q]}}{\binom{W}{M}}. \tag{20}$$

To get the desired probability of obtaining all U primary classes in N trials from the W secondary classes, we multiply the probabilities given by Equations (14) and (17) and sum over all appropriate cases:

$$P(N,U,m_1,m_2,\ldots m_u) = \sum_{j=0}^{\min(N,W)-U} P(N,W-M)\, Z(U,U+j,W), \tag{21}$$

where M = W - (U+j).

These results should be considered in light of well-known results for the probabilities of obtaining specific numbers of items from each category. For sampling without replacement or for sampling from an infinite population, the multinomial distribution applies. The probability of getting $x_1, x_2, \ldots x_u$ items from U different categories through random sampling with replacement (or from an infinite population) is[4]

$$P(x_1, x_2,\ldots x_u) = \frac{N!}{x_1!\, x_2!\ldots x_u!}\, p_1^{x_1}\, p_2^{x_2}\ldots p_u^{x_u} \tag{22}$$

where p_i is the probability of sampling an item from category U_i, and the total sample size is $N = \Sigma x_i$. A different form for the probability given by Equation (21) can be obtained by summing $P(x_1, x_2,\ldots x_u)$ of Equation (22) for all combinations of $x_1, x_2,\ldots x_u$ with $x_i \neq 0$ and $\Sigma x_i = N$. This approach may often be more convenient than applying Equation (21).

The hypergeometric distribution applies when the sampling is done without replacement from a finite population. The number of items in any category, U_i, of the U categories is m_i. The total number of items in the population is $W = \sum m_i$. N of the W items will be sampled, with $N = \sum x_i$ as before. The probability of obtaining a specific configuration is given by[5]

$$P(x_1, x_2, \ldots x_u) = \frac{\binom{m_1}{x_1}\binom{m_2}{x_2}\cdots\binom{m_u}{x_u}}{\binom{W}{N}} . \qquad (23)$$

FURTHER PROPERTIES OF THE TWO-DIMENSIONAL FACTORIAL

The two-dimensional factorial appears to be an interesting function meriting further study. Table 2 shows values of F(D,U-M) for $1 \le D \le 25$ and $1 \le U\text{-}M \le 7$. Several interesting features are apparent in the columns of numbers shown here. Note that F(D,1) = 1 and $F(D,2) = 2^{D+1} - 1$ for all D. A logarithmic contour plot in Figure 1 for the range $1 \le D \le 30$ and $1 \le U\text{-}M \le 29$ shows how the numbers increase with U and D.

D \ U-M	1	2	3	4	5	6	7
1	1	3	6	10	15	21	28
2	1	7	25	65	140	266	462
3	1	15	90	350	1050	2646	5880
4	1	31	301	1701	6951	22827	63987
5	1	63	966	7770	42525	179487	627396
6	1	127	3025	34105	246730	1323652	5715424
7	1	255	9330	145750	1379400	9321312	49329280
8	1	511	28501	611501	7508501	63436373	408741333
9	1	1023	86526	2532530	40075035	420693273	3281882604
10	1	2047	261625	10391745	210766920	2734926558	25708104786
11	1	4095	788970	42355950	1096190550	17505749898	1.97E+11
12	1	8191	2375101	171798901	5652751651	110687251039	1.49E+12
13	1	16383	7141686	694337290	28958095545	6.93E+11	1.11E+13
14	1	32767	21457825	2798806985	147589284710	4.31E+12	8.23E+13
15	1	65535	64439010	11259666950	7.49E+11	2.66E+13	6.03E+14
16	1	131071	193448101	45232115901	3.79E+12	1.63E+14	4.38E+15
17	1	262143	580606446	1.82E+11	1.91E+13	9.99E+14	3.17E+16
18	1	524287	1742343625	7.28E+11	9.64E+13	6.09E+15	2.28E+17
19	1	1048575	5228079450	2.92E+12	4.85E+14	3.70E+16	1.63E+18
20	1	2097151	15686335501	1.17E+13	2.44E+15	2.25E+17	1.16E+19
21	1	4194303	47063200806	4.68E+13	1.22E+16	1.36E+18	8.29E+19
22	1	8388607	1.41E+11	1.87E+14	6.13E+16	8.22E+18	5.88E+20
23	1	16777215	4.24E+11	7.49E+14	3.07E+17	4.96E+19	4.17E+21
24	1	33554431	1.27E+12	3.00E+15	1.54E+18	2.99E+20	2.95E+22
25	1	67108863	3.81E+12	1.20E+16	7.71E+18	1.80E+21	2.08E+23

Table 2. F(D,U-M) for $1 \le D \le 25$ and $1 \le U\text{-}M \le 7$.

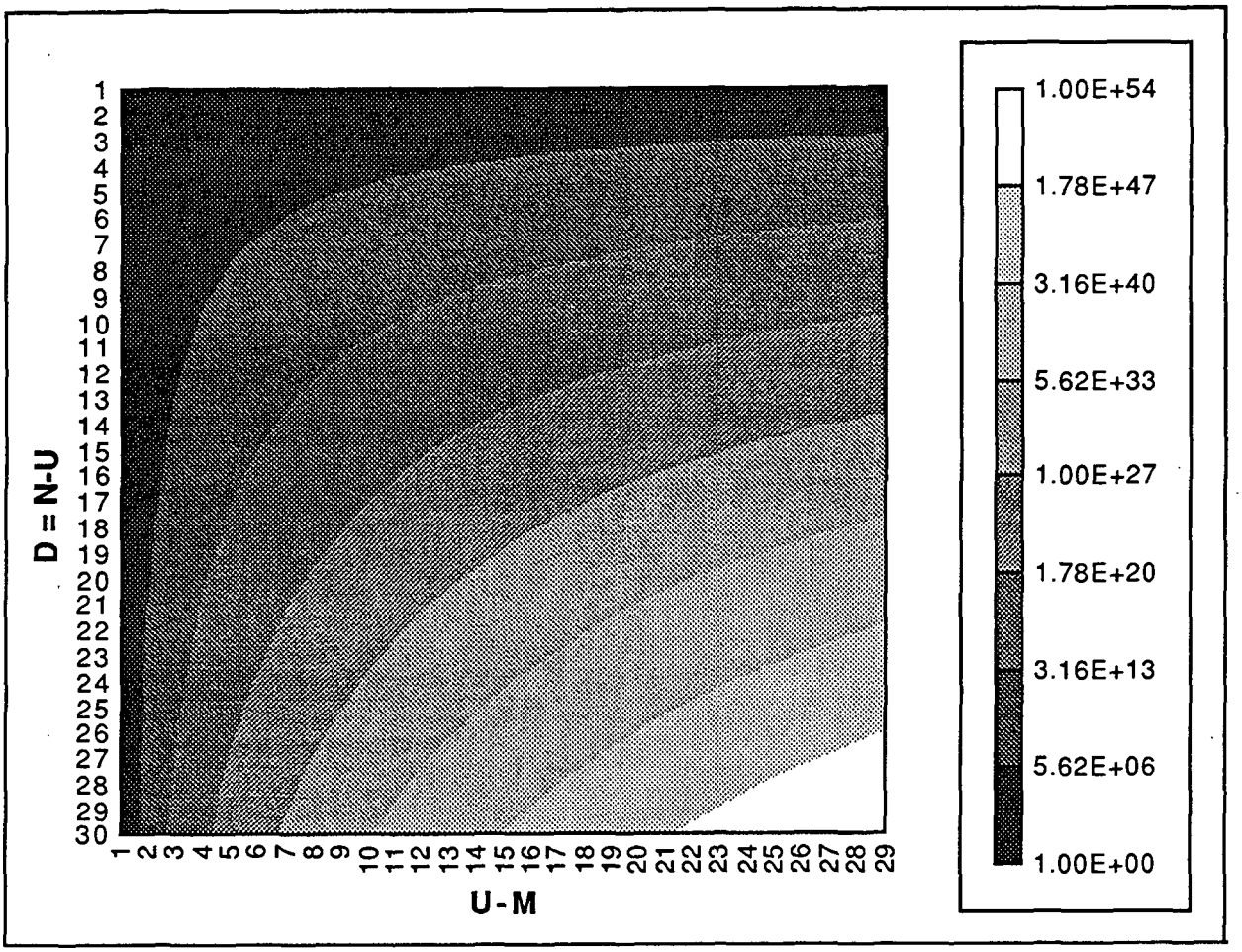


Figure 1. Logarithmic contour plot of F(D,U-M) for $1 \le D \le 30$ and $1 \le U-M \le 29$.

Limits for Large Numbers

As D, and hence N, becomes very large for a given U, P(N,U) approaches unity (it becomes nearly certain that all U items will be collected if enough samples are obtained). Thus,

$$\lim_{D\to\infty} F(D,U) = \frac{U^N}{U!}. \tag{24}$$

Therefore, the ratio F(D,U)/F(D-1,U) approaches U for large D. Likewise, for large D, the ratio of adjacent values in any row is

$$\lim_{D\to\infty} \frac{F(D,U)}{F(D,U-1)} = \left(\frac{U}{U-1}\right)^{N-1} \tag{25}$$

A more exact expression than Equation (24) is possible using a theorem for the limit of Equation (1) proved by Feller[6] and attributed (with a different proof) to von Mises[7]:

If U and N increase so that $\lambda = Ue^{-N/U}$ *remains bounded, then for fixed M:*

$$P(N,U-M) \rightarrow \frac{\lambda^M}{M!} e^{-\lambda}, \tag{26}$$

which is the Poisson distribution. The two-dimensional factorial for large N = U+D is then

$$F(D,U-M) = F(N-U+M,U-M) \rightarrow \frac{U^{(N+M)}}{U!} \exp -\left[\frac{NM}{U} + U \exp\left(\frac{-N}{U}\right)\right]. \tag{27}$$

For M = 0, this can be rewritten as

$$F(N-U,U) \rightarrow \frac{\left(U^U\right)^{N/U} \left(e^{-U}\right)^{e^{-N/U}}}{U!}, \tag{28}$$

or for finite M, we can re-express Equation (26) as

$$F(D,U-M) = F(N-U+M,U-M) \rightarrow \frac{\left(M^M\right)^{\ln(M/U)} \left(e^{-M}\right)^{N/U} \left(U^U\right)^{N/U} \left(e^{-U}\right)^{e^{-N/U}}}{U!}, \tag{29}$$

where the terms in the numerator bear some resemblance to Stirling's formula for large factorials,

$$n! \rightarrow \sqrt{2\pi n}\, n^n\, e^{-n}. \tag{30}$$

While the resemblance to the regular factorial function is somewhat superficial, the two-dimensional factorial is still suggested as an appropriate name for the recursive F function introduced here. The main similarity to the factorial is through the recursive expression given in Equation (12).

Comparisons of the approximate form in Equation (26) with the exact probability form in Equation (11) suggest that the approximate form must be used with caution for M>0. For example, for a given U and M, the approximation may be close for a certain range of N, but will become increasingly incorrect as N increases.

Number Analysis

One feature of the numbers produced by the two-dimensional factorial is that a large proportion of them seem to have seven and eleven as factors. In Table 3 (a subset of Table 2), numbers divisible by seven are in italics, and numbers divisible by eleven are in boldface. About 20% of the numbers examined are divisible by both seven and eleven. I have no explanation for this feature.

D \ U-M	1	2	3	4	5	6	7
1	1	3	6	10	15	*21*	*28*
2	1	*7*	25	65	*140*	*266*	***462***
3	1	15	90	*350*	*1050*	*2646*	*5880*
4	1	31	*301*	*1701*	*6951*	*22827*	***63987***
5	1	*63*	*966*	*7770*	*42525*	***179487***	***627396***
6	1	127	**3025**	34105	**246730**	**1323652**	**5715424**
7	1	255	9330	**145750**	**1379400**	***9321312***	***49329280***
8	1	*511*	**28501**	**611501**	***7508501***	***63436373***	***408741333***
9	1	**1023**	**86526**	***2532530***	***40075035***	***420693273***	***3281882604***

Table 3. Subset of Table 2 showing F(D,U-M) values divisible by seven in italic and values divisible by eleven are in boldface.

Examination of the last digits of the numbers in columns 2 through 5 shows interesting repeating patterns if we consider that the initial, undisplayed row for D = 0 consists of ones. The repeating final digits are:

Column 2: 1-3-7-5
Column 3: 1-6-5-0
Column 4: 1-0-5-0
Column 5: 1-5-0-0

Column 6 shows an interesting pattern in the final digits. The sequence is 1-1 – 6-6 – 7-7 – 2-2 – 3-3 – 8-8 – 9-9 – 4-4 – 5-5 – 0-0, which apparently repeats (I am not sure because of limited numerical resolution). These pairs of digits change according to a specific pattern: add 5, add 1, subtract 5, add 1, and repeat.

APPLICATIONS AND EXAMPLES

Probability of Collecting *at Least* U-M Sets

P(N,U-M) in Equation (11) gives the probability of obtaining exactly U-M identical sets in a random sample of size N from a uniform, infinite population. The collector, however, is usually more interested in the probability of collecting at least a specified number of distinct items. For varying M with constant N and U, each P(N,U-M) is independent. Therefore, the probability that no more than M_{max} classes are missing in a random sample of size N is given by

$$P(N,U\text{-}[M\le M_{max}]) = \frac{U!}{U^N}\sum_{M=0}^{M_{max}} \frac{F(N\text{-}U+M,U\text{-}M)}{M!}. \qquad (31)$$

Since the probability of having at least one unique item is unity,

$$\frac{U!}{U^N}\sum_{M=0}^{N-U} \frac{F(N\text{-}U+M,U\text{-}M)}{M!} = 1. \qquad (32)$$

Expected Number of Trials to Complete a Set

For a series of U distinct items sampled with replacement, Feller[8] shows that the expected number of samples to obtain U-M distinct items is

$$E(N_{U-M}) = U\left\{\frac{1}{U} + \frac{1}{U-1} + \frac{1}{U-2} + \ldots + \frac{1}{M+1}\right\}, \tag{33}$$

which, for large U, can be approximated by

$$E(N_{U-M}) \approx U \ln\left(\frac{U}{M+1}\right). \tag{34}$$

Equation (11) can be applied to determine the probability that U-M items are indeed collected in $E(N_{U-M})$ samples. A plot of $P(E(N_{U-M}),U-M)$ versus U is given for M = 0 in Figure 2, using $E(N_{U-M})$ values rounded up to the next highest integer.

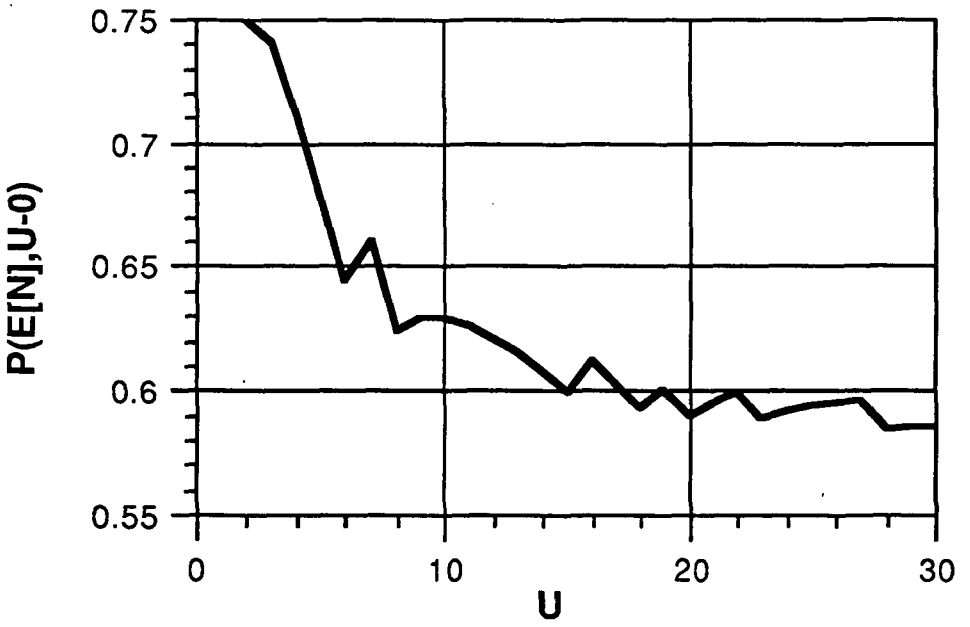


Figure 2. Plot of P(E[N],U), the probability of obtaining all U sets in the expected number of trials (Equation [31]), versus U.

In the limit of large U, E(N) = U ln(U) when M = 0. Applying the Poisson approximation to P(N,U-0) for large N, we see that the probability of collecting all U sets in E(N) trials approaches $e^{-1} \approx 0.3679$ as U becomes large. In Figure 2, this limit is still far off at U = 30.

Sample Probability Results

For M = 0, a table of P(N,U) values from Equation (11) for uniform distributions is given in Table 4, where the numbered columns correspond to U and the numbered rows to D = N-U. This investigation began with a consideration

D = N - U \ U	1	2	3	4	5	6	7	8	9	10	11	12
1	1.0000	0.7500	0.4444	0.2344	0.1152	0.0540	0.0245	0.0108	0.0047	0.0020	0.0008	0.0003
2	1.0000	0.8750	0.6173	0.3809	0.2150	0.1140	0.0577	0.0282	0.0134	0.0062	0.0028	0.0013
3	1.0000	0.9375	0.7407	0.5127	0.3226	0.1890	0.1049	0.0558	0.0286	0.0143	0.0069	0.0033
4	1.0000	0.9688	0.8258	0.6229	0.4271	0.2718	0.1631	0.0933	0.0513	0.0273	0.0141	0.0071
5	1.0000	0.9844	0.8834	0.7114	0.5225	0.3562	0.2285	0.1393	0.0815	0.0460	0.0251	0.0134
6	1.0000	0.9922	0.9221	0.7806	0.6064	0.4378	0.2973	0.1917	0.1183	0.0703	0.0404	0.0226
7	1.0000	0.9961	0.9480	0.8340	0.6780	0.5139	0.3666	0.2482	0.1607	0.1001	0.0602	0.0352
8	1.0000	0.9980	0.9653	0.8748	0.7381	0.5828	0.4339	0.3068	0.2073	0.1347	0.0845	0.0514
9	1.0000	0.9990	0.9769	0.9057	0.7879	0.6442	0.4977	0.3656	0.2567	0.1732	0.1128	0.0711
10	1.0000	0.9995	0.9846	0.9291	0.8288	0.6980	0.5570	0.4231	0.3075	0.2147	0.1447	0.0944
11	1.0000	0.9998	0.9897	0.9467	0.8621	0.7446	0.6112	0.4783	0.3585	0.2583	0.1796	0.1209
12	1.0000	0.9999	0.9931	0.9600	0.8891	0.7847	0.6601	0.5306	0.4088	0.3031	0.2168	0.1502
13	1.0000	0.9999	0.9954	0.9700	0.9109	0.8189	0.7039	0.5793	0.4577	0.3481	0.2557	0.1819
14	1.0000	1.0000	0.9970	0.9775	0.9286	0.8480	0.7427	0.6243	0.5044	0.3928	0.2956	0.2155
15	1.0000	1.0000	0.9980	0.9831	0.9427	0.8726	0.7770	0.6654	0.5487	0.4366	0.3359	0.2504
16	1.0000	1.0000	0.9986	0.9873	0.9541	0.8933	0.8071	0.7028	0.5904	0.4790	0.3761	0.2863
17	1.0000	1.0000	0.9991	0.9905	0.9632	0.9108	0.8334	0.7366	0.6291	0.5196	0.4157	0.3227
18	1.0000	1.0000	0.9994	0.9929	0.9706	0.9254	0.8562	0.7670	0.6650	0.5583	0.4544	0.3591
19	1.0000	1.0000	0.9996	0.9946	0.9764	0.9377	0.8761	0.7943	0.6980	0.5948	0.4919	0.3953
20	1.0000	1.0000	0.9997	0.9960	0.9811	0.9480	0.8933	0.8185	0.7283	0.6291	0.5280	0.4309
21	1.0000	1.0000	0.9998	0.9970	0.9849	0.9566	0.9082	0.8401	0.7559	0.6612	0.5624	0.4656
22	1.0000	1.0000	0.9999	0.9977	0.9879	0.9638	0.9211	0.8593	0.7810	0.6910	0.5952	0.4993
23	1.0000	1.0000	0.9999	0.9983	0.9903	0.9698	0.9322	0.8763	0.8037	0.7186	0.6261	0.5318
24	1.0000	1.0000	0.9999	0.9987	0.9923	0.9748	0.9418	0.8913	0.8243	0.7440	0.6552	0.5630
25	1.0000	1.0000	1.0000	0.9990	0.9938	0.9790	0.9500	0.9045	0.8428	0.7675	0.6825	0.5928
26	1.0000	1.0000	1.0000	0.9993	0.9950	0.9825	0.9571	0.9162	0.8595	0.7890	0.7081	0.6211
27	1.0000	1.0000	1.0000	0.9995	0.9960	0.9854	0.9632	0.9265	0.8746	0.8087	0.7318	0.6479
28	1.0000	1.0000	1.0000	0.9996	0.9968	0.9878	0.9684	0.9355	0.8880	0.8267	0.7539	0.6732
29	1.0000	1.0000	1.0000	0.9997	0.9975	0.9899	0.9729	0.9435	0.9001	0.8431	0.7744	0.6971
30	1.0000	1.0000	1.0000	0.9998	0.9980	0.9915	0.9767	0.9505	0.9109	0.8581	0.7934	0.7195
31	1.0000	1.0000	1.0000	0.9998	0.9984	0.9930	0.9801	0.9566	0.9206	0.8717	0.8109	0.7404
32	1.0000	1.0000	1.0000	0.9999	0.9987	0.9941	0.9829	0.9620	0.9293	0.8841	0.8271	0.7601
33	1.0000	1.0000	1.0000	0.9999	0.9990	0.9951	0.9853	0.9667	0.9370	0.8953	0.8419	0.7784
34	1.0000	1.0000	1.0000	0.9999	0.9992	0.9959	0.9874	0.9708	0.9439	0.9055	0.8556	0.7954
35	1.0000	1.0000	1.0000	0.9999	0.9993	0.9966	0.9892	0.9744	0.9500	0.9147	0.8681	0.8113
36	1.0000	1.0000	1.0000	1.0000	0.9995	0.9972	0.9908	0.9776	0.9555	0.9230	0.8796	0.8260
37	1.0000	1.0000	1.0000	1.0000	0.9996	0.9976	0.9921	0.9804	0.9604	0.9306	0.8902	0.8397
38	1.0000	1.0000	1.0000	1.0000	0.9997	0.9980	0.9932	0.9829	0.9648	0.9374	0.8998	0.8523
39	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9942	0.9850	0.9687	0.9435	0.9087	0.8641
40	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986	0.9950	0.9869	0.9721	0.9491	0.9168	0.8749

Table 4. P(N,U) values for M = 0. Column numbers correspond to U and row numbers to D = N-U.

of the difficulty of collecting complete sets of items offered randomly inside cereal boxes, and we will briefly discuss this problem here. Based on Table 4, the would-be collector should plan on buying three to five times as many packages as there are items to be collected to be fairly sure (ca. 90% confident) of collecting a complete set with less than 20 items. For larger sets (say > 25 items), it may be necessary to buy six or more times as many packages as there are items to be collected. In the case of U = 50 license plates, Table 5 shows the probabilities of completing sets with various M values if one buys N = 100 boxes. The likelihood of collecting plates from all 50 states is 0.00017, and the chance that no more than three states will be missing is only 5.18%. The most likely outcome is that six states will be missing, although there is a 52% probability that even more than six will be missing. With N = 180, the probability of completing the set is still only 24.5% (25.5% according to the approximation of Equation (26). To be 90% confident of getting all 50 states, an estimated 308 boxes must be purchased. (Consumers may do well to simply contact the manufacturer of the collectable items and buy a complete set directly.)

M	P(100,50-M)	Cumulative
0	0.00017	0.00017
1	0.00202	0.00219
2	0.01129	0.01348
3	0.03835	0.05183
4	0.08910	0.14093
5	0.15071	0.29164
6	0.19294	0.48458
7	0.19183	0.67641
8	0.15082	0.82723
9	0.09501	0.92224
10	0.04841	0.97065
11	0.02009	0.99074
12	0.00682	0.99756
13	0.00190	0.99947
14	0.00044	0.99990
15	0.00008	0.99999

Table 5. Probabilities for the case of U = 50 and N =100.

CLOSURE

A new form of the solution to a classical probability problem has yielded an interesting function which may be termed a two-dimensional factorial. The function allows computation of set collection probabilities with improved accuracy compared to the classical alternating-sign series solution in Equation (1) for uniformly distributed populations and is also of use in the more complex case of nonuniform populations.

ACKNOWLEDGMENT

The author is grateful for the valuable comments and suggestions offered by Dr. Bruce Collings of the Brigham Young University Statistics Department and by Kendra L. Lindsay, the author's spouse and statistical consultant.

REFERENCES

1. Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley and Sons, New York, NY (1950), pp. 51-66.

2. Feller, p. 69, see also p. 64.

3. Feller, pp. 60-61.

4. Gunther, W. C., *Concepts of Probability*, McGraw-Hill, New York (1968), p. 172.

5. Gunther, pp. 180-181.

6. Feller, pp. 72-75.

7. von Mises, R., "Über Aufteilungs- und Besetzungswarhscheinlichkeiten," *Revue de la Faculté des Sciences de l'Université d'Istanbul*, N.S., Vol. 4 (1939), pp. 1-19, as cited by Feller, p. 72.

8. Feller, pp. 174-175.